



ENSEIRB-MATMECA

FILIÈRE INFORMATIQUE, 2<sup>ème</sup> ANNÉE

---

## Rapport de Stage Industriel 2A

---

*Auteur :*  
LUX BENJAMIN

*Encadrant :*  
LAPOIRE DENIS

*Stagiaire :*  
LUX BENJAMIN

*Maitre de stage :*  
GOLHEN RONAN

26 novembre 2012

## Table des matières

<b>1</b>	<b>Abstract</b>	<b>2</b>
<b>2</b>	<b>Présentation du stage</b>	<b>2</b>
2.1	Présentation de l'entreprise <i>Maxsea</i>	2
2.2	Contacts : <i>Maxsea</i>	2
2.3	Description du stage	2
2.4	Acteurs du stage	3
2.5	Description de l'existant	3
2.6	Mission Technique	4
<b>3</b>	<b>Réalisation du stage</b>	<b>4</b>
3.1	Lignes de conduites	4
3.2	Procédure de génération de cartes	4
3.3	Services fournis par la plateforme <i>Windows Azure</i>	5
3.3.1	Déploiement Azure	5
3.3.2	Worker Roles	5
3.3.3	Azure Storage	6
3.3.4	Azure SQL	6
3.4	Détail de l'architecture et choix d'implémentation	6
3.4.1	La classe <i>WorkerRole</i>	7
3.4.2	La classe <i>Worker</i>	7
3.4.3	Les classes héritant de <i>BusinessProcess</i>	7
3.4.4	La classe <i>Manager</i>	8
3.5	Détail des tâches métiers	9
3.5.1	La classe <i>Reprojection</i>	9
3.5.2	La classe <i>MainProduction</i>	9
3.5.3	La classe <i>PostProduction1</i>	9
3.5.4	La classe <i>PostProduction2</i>	10
3.5.5	La classe <i>Copie</i>	10
3.6	Résultats	10
3.6.1	Fichier au format dbv	10
3.6.2	Péformance de travail	10
3.6.3	Budgétaires	11
<b>4</b>	<b>Amélioration du prototype</b>	<b>11</b>
4.1	Todo-List	11
4.2	Conclusion	11

## 1 Abstract

Ce document constitue le rapport de stage de Benjamin Lux dans l'entreprise *Maxsea*. Élève en 2<sup>ème</sup> année, en section informatique, à l'école d'ingénieur ENSEIRB-MATMECA, à BORDEAUX.

## 2 Présentation du stage

### 2.1 Présentation de l'entreprise *Maxsea*

*Maxsea International* est une entreprise (SAS) de production de logiciels maritimes. Son produit phare "Maxsea Time Zero" est un logiciel d'aide à la navigation maritime. Les bâtiments de son siège social, à Bidart, sur la côte Basque, habritent une soixantaine d'employés. La moitié sont des développeurs qui maintiennent le logiciel, l'autre fait partie de la filiale *Map-Media* qui produit les cartes lisibles par le logiciel, ces derniers n'ont donc pas de formation de développeurs. Il existe deux autres filiales, une en Espagne pour les relations commerciales et une de développeurs aux États-Unis.

### 2.2 Contacts : *Maxsea*

- Adresse : Technopole Izarbel, 64210 Bidart – France
- mail : info@maxsea.fr
- tel : + 33 559 43 81 00
- fax : + 33 559 43 81 01

### 2.3 Description du stage

**Titre :**

Migration d'une chaîne de production cartographique vers un Cloud.

**Mots-clés :**

Cloud computing, Production cartographique, Windows Azure, Géomatique, Calcul parallèle.

**Sujet du stage :**

Dans le cadre de la production de données cartographiques vecteur, la société *Maxsea* veut employer les technologies "Cloud" et en particulier la solution "Azure" de *Microsoft*. Pour cela, il s'agira de coder des logiciels en *C#*, s'appuyant sur le Framework open source *GDal/OGR*<sup>1</sup> (*OGR* pour le traitement de données vecteur). Ceux-ci traiteront des étapes de la production de données. Leur intérêt est la forte montée en charge possible dans le Cloud, afin de réduire les temps de production (2 lots de production par an, couvrant l'ensemble des cartes marines du globe).

**Dates de stage :**

Le stage commence le 6 juin 2012 et finit le 21 septembre 2012.

---

1. *Geospatial Data Abstraction Library* license by the *Open Source Geospatial Foundation* (*OSGeo*)

## 2.4 Acteurs du stage

### Maitre de Stage : Ronan GOLHEN

- tel : + 33 614 17 65 05
- mail : ronan.golhen@maxsea.fr

### Tuteurs de stages

#### Arnaud CAPDEVIELLE

- mail : arnaud.capdevielle@maxsea.fr

#### Arnaud REMY

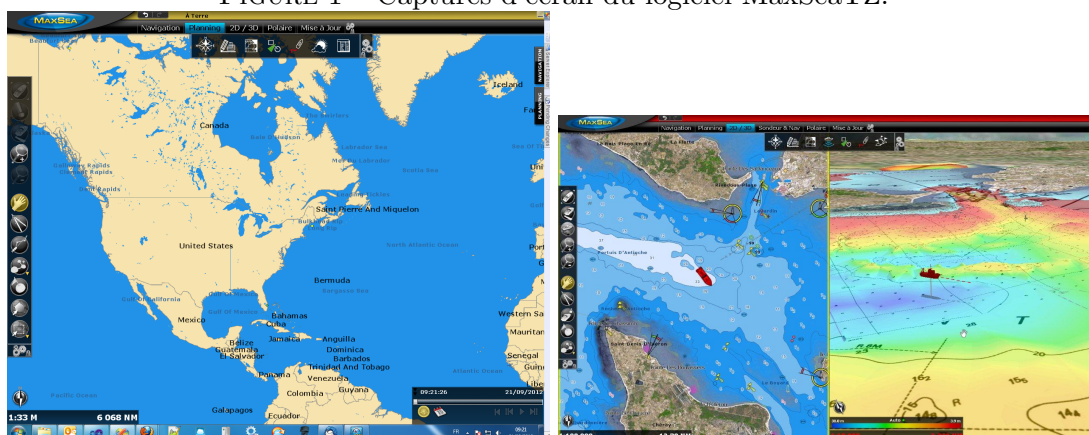
- mail : arnaud.remy@maxsea.fr

## 2.5 Description de l'existant

Pour produire ses cartes, la société *Maxsea* reçoit chaque année diverses livraisons de cartes au format S57. Ses deux principaux fournisseurs, *Jeppesen* et *Navionics*, livrent deux fois par an l'ensemble des cartes maritimes du monde. Une seule de ces livraisons contient 45000 fichiers. Au début du stage la chaîne de production des cartes au format dbv était localisé dans les locaux de l'entreprise. Le temps de transformation était cependant préoccupant, puisque les livraisons prenaient du retard. En effet, en utilisant les 12 serveurs disponibles pour cette opération de production, quatre mois étaient nécessaires au traitement d'une livraison.

La chaîne de production de cartes vecteurs est supervisée par le logiciel SSIS<sup>2</sup> de *Microsoft*. Une fois les données reçues, par DVD ou FTP<sup>3</sup>, un opérateur lance l'exécution de l'ETL<sup>4</sup> SSIS. Celui ci va lancer un par un les différents traitements qui seront nécessaires à la production. L'opérateur effectue différents contrôles plus ou moins automatisés, notamment à la fin de la création du catalogue. Durant ce processus les cartes sont manipulées sous forme de bases de données spatiales. Les deux SGBD<sup>5</sup> utilisées pour ces traitements sont SQL Server et PostgreSQL. Les opérations sur celles-ci se font principalement via FME<sup>6</sup>, un ETL Spatial, qui s'appuie sur le Framework open source spatial *GDal/OGR*.

FIGURE 1 – Captures d'écran du logiciel MaxSeaTZ.



2. SQL Server Integration Services
3. *File Transfert Protocol*, Protocole de transfert de fichiers
4. *Extract Transform Load*, extraction transformation alimentation
5. *Système de Gestion de Base de Données*
6. Logiciel propriétaire de SAFE SOFTWARE

## 2.6 Mission Technique

Le stage consiste à déployer la chaîne de production dans un environnement de cloud computing. Ceci doit amener trois avantages à la société. Le premier et le plus important est de pouvoir produire les cartes dans le temps demandé. Ensuite vient la libération de ressources machine au sein des locaux et la baisse du prix de revient de la chaîne de production vecteur.

## 3 Réalisation du stage

### 3.1 Lignes de conduites

Les pistes explorées par Ronan GOLHEN mettent en avant l'offre de *PaaS*<sup>7</sup> Azure de *Microsoft*. Les caractéristiques principales de cette offre sont de libérer la société de la gestion des serveurs et de ne payer qu'à l'usage. Sa philosophie incite fortement aux modèles de calcul parallèle.

Une architecture logicielle permettant de traiter une partie ou l'ensemble de la production vecteur à l'intérieur du Cloud *Microsoft* doit être mise en place. Ce travail devra être exploitable par les opérateurs de *MapMedia* sans apprentissage des technologies de développement.

Migrer la chaîne de production demande une bonne connaissance des principes de workflow, de développement Azure et du Framework métier *OGR*.

Le workflow est à reconsidérer, on ne peut pas juste copier la chaîne existante pour deux raisons principales.

Tout d'abord le service hébergeur est distant et on ne peut pas se permettre de contrôler toute la production dans les locaux, ce qui impliquerait un rapatriement des données entre chacune de ces étapes (plusieurs Giga-octets).

Ensuite le fait que la chaîne utilise des logiciels propriétaires pose un problème de licence car ils sont liés à une machine physique, principe à bannir en terme de technologie cloud. Nous avons contacté SAFE SOFTWARE, la société qui développe FME, cependant celle-ci n'est pas prête à fournir des licences payantes au nombre d'utilisateurs. Leur principe de licence prévoit que leur logiciel doit s'exécuter toujours sur la même machine, chose inenvisageable en univers cloud.

Azure possède un Framework particulier qu'il faut maîtriser. Par exemple les espaces de stockage permanent sont séparés des ressources physiques de la machine, qui elles, deviennent des ressources temporaires. Ainsi à chaque étape de calcul on sauvegarde les résultats dans des espaces de stockage persistant appelé Azure Blob.

*OGR* sera la principale alternative à FME. En effet on doit pouvoir se passer de ce logiciel propriétaire et on peut se servir du Framework qu'il utilise puisque ce dernier est open source. Il a donc fallu recoder les étapes de production sans passer par l'intermédiaire de FME.

### 3.2 Procédure de génération de cartes

Pour produire les cartes au format dbv, l'opérateur utilisant la solution créée pendant le stage doit faire quelques manipulations.

Il doit commencer par générer des tables SQL avec l'ancien système (CATALOGUE\_RCL\_ALL, MX\_CHART, ...). Il les migre ensuite sur un serveur *SQL Azure*, grâce au logiciel *SQL Database*

---

7. *Platform as a Service*, Plateforme en tant que service

*Migration Wizard*<sup>8</sup>. De même grâce à des logiciels comme *Azure Storage Exploreur* ou *Cloud-Berry* les datas d'entrée, les cartes au format .s57, sont placées dans des Blobs Azure.

Après avoir préparé un espace de travail azure, il déploie la solution sur le Cloud Azure. Une fois les instances de travail opérationnelles, il peut utiliser les différents scripts, fichiers .bat, mis à sa disposition pour exécuter les différentes étapes de la création du dbv. Dans l'ordre ces étapes se nomment : Reprojection, MainProduction, PostProduction1, PostProduction2 et Copie.

Enfin, il récupère les résultats obtenus (fichier dbv) à partir d'un blob azure.

### 3.3 Services fournis par la plateforme *Windows Azure*

Cette partie a pour but de familiariser le lecteur avec les concepts Azure que nous emploierons plus tard. Nous allons évoquer des services proposés par Microsoft.

#### 3.3.1 Déploiement Azure

L'offre de *PaaS* Azure s'appuie principalement sur le fait de déployer facilement des logiciels. Pour ce faire il faut développer un code qui sera zippé dans un fichier de binaire et ressources (extension .cspkg) et un fichier de configuration (extension .cscfg). Une fois ces deux fichiers uploadés, via le portail Azure<sup>9</sup>, on peut gérer notre déploiement.

#### 3.3.2 Worker Roles

La cellule de base de l'architecture logicielle apportée est le Worker Role, autrement dit le rôle de travail.

Lorsqu'on parle de worker role on fait souvent référence au code qui le compose. L'unité de travail dans le déploiement Azure, qui correspond à ce worker role, est une instance de ce worker role. On peut faire ici un parallèle avec un objet et une classe en POO<sup>10</sup>.

Une implémentation de cette fonctionnalité est fournie par l'API<sup>11</sup>. Il s'agit d'une classe, C# dans notre cas, nommée RoleEntryPoint. On peut créer une classe héritant de celle-ci et redéfinir les méthodes *OnStart()*, *Run()* et *OnStop()*. La méthode la plus importante est la méthode *Run()*, on y met le code que l'on veut exécuter.

Un worker role est censé tourner en boucle, ceci sans jamais s'arrêter. Le processus peut tout de même s'interrompre dans des cas particuliers. Si l'utilisateur décide de stopper son déploiement ou si le rôle est transféré sur une autre machine alors la méthode *OnStop()* est appelée. Dans le cas d'un bug, bug du programme comme la monopolisation de toutes les ressources ou bug externe comme une coupure de courant, la méthode *OnStop()* ne peut être appelée.

---

8. Logiciel gratuit de migration de base de donnée, disponible sur [sqlazuremw.codeplex.com](http://sqlazuremw.codeplex.com)

9. <https://manage.windowsazure.com/>

10. Programmation Orientée Objet

11. <http://msdn.microsoft.com/en-us/library/microsoft.windowsazure.serviceruntime.roleentrypoint.members.aspx>

Le processus correspondant à une instance d'un worker role est exécuté sur une VM<sup>12</sup> spécialement montée pour lui. On dispose sur cette machine, en fonction des options de configuration, d'un certain espace mémoire et d'une certaine quantité de disque. Les données enregistrées sur le disque sont effacées avec l'interruption de la VM, lors de l'arrêt de l'instance ou d'un dysfonctionnement du serveur. De plus ces données sont local à l'instance, les autres instances ne peuvent y accéder.

Pour pouvoir sauvegarder des résultats accessibles par tous, autres instances d'un même déploiement et utilisateurs extérieurs, la plateforme Azure propose deux outils, l'Azure Storage et Azure SQL. On peut bien sûr coder nos propres outils, qui upload les données des instances sur un serveur de l'entreprise par exemple. Cependant ces outils ont deux avantages majeurs. Ils sont déjà faits et sont situés sur des machines proches des VM, d'un point de vue géographique, ce qui augmente considérablement la bande passante.

### 3.3.3 Azure Storage

Azure Storage est un espace de stockage de données numériques de grande capacité. On y a trois services différents, les blobs, les queues et les tables.

Les blobs sont des objets binaires non structurés, on peut les assimiler à des fichiers. Ils sont regroupés dans des conteneurs. On peut assimiler les blobs à des fichiers et les conteneurs à des dossiers dans un système de fichiers ; à la différence qu'un conteneur ne peut pas posséder de conteneur. Les queues sont des files FIFO<sup>13</sup> de messages, texte ou binaires. Elles implémentent plusieurs mécanismes permettant de garantir la pérennité et la survie des messages. Les tables sont des conteneurs de données structurées, elles peuvent contenir de nombreux objets de petite taille.

### 3.3.4 Azure SQL

Azure SQL est un serveur SQL hébergé dans des machines proches des VM. Il s'agit en réalité d'un SQL Server légèrement modifié.

## 3.4 Détail de l'architecture et choix d'implémentation

Pour accompagner la construction du projet au niveau architecture logicielle deux réunions ont été organisées avec des participants extérieurs à *Maxsea*. La première avec des responsables du *MCIA*<sup>14</sup> a surtout abouti sur des considérations théoriques sur le cloud computing. Ces personnes ont davantage l'habitude de travailler sur des offres *IaaS* et sur des environnements UNIX. La seconde fut à Paris avec un expert en architecture Azure de Microsoft France et deux employés de la société YSANCE, qui ont fait un projet semblable au notre. Cette dernière fut plus concrète car les participants avaient l'habitude de manipuler les API de Windows Azure.

La solution apportée consiste en un déploiement et donc en une solution C# Visual Studio<sup>15</sup>.

---

12. *Virtual Machine*, Machine Virtuelle

13. First In First Out, Premier entré premier sorti

14. *Mésocentre de Calcul Intensif Aquitain*

15. *IDE* (Integrated development Environment) de Microsoft.

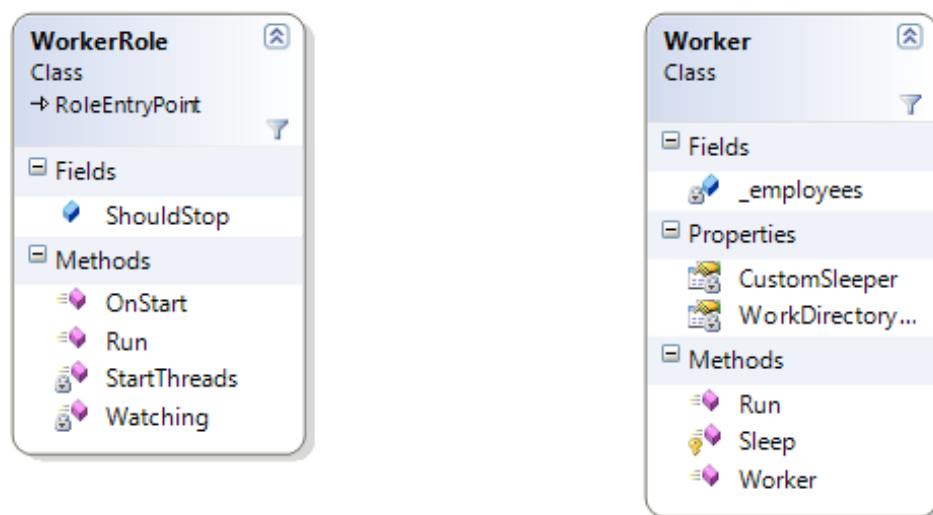
### 3.4.1 La classe *WorkerRole*

La classe *Worker Role* constitue le point d'entrée du programme, elle hérite d'ailleurs de *RoleEntryPoint*. Celle ci va instancier n threads de travail, puis va entrer dans une boucle infinie de surveillance de ces threads. La création de ces threads ce fait grace à la classe *Thread*. Ils instancient une classe *Worker*. La surveillance n'a pas était implémentée, elle consiste à vérifier qu'il n'y a pas de fuite mémoire, pour au besoin stopper ou relancer des threads de travail.

### 3.4.2 La classe *Worker*

La classe *textitWorker* sert à trouver et répartir les tâches de travail. Dans son instantiation elle crée dynamiquement, en utilisant le ..... , un tableau de d'objet héritant de la classe abstraite *BusinessProcess* et implémentant l'interface *IBusinessProcess*. Cette interface demande d'implémenter deux méthodes, à savoir *Init()* et *Work()*. La seconde renvoie *false* si l'objet ne trouve pas de travail et *true* après avoir traité une tâche sinon. L'objet *Worker* bouclera sur le tableau pour que toutes les tâches soit traitées.

FIGURE 2 – Diagramme de classe partiel du Worker Role.



### 3.4.3 Les classes héritant de *BusinessProcess*

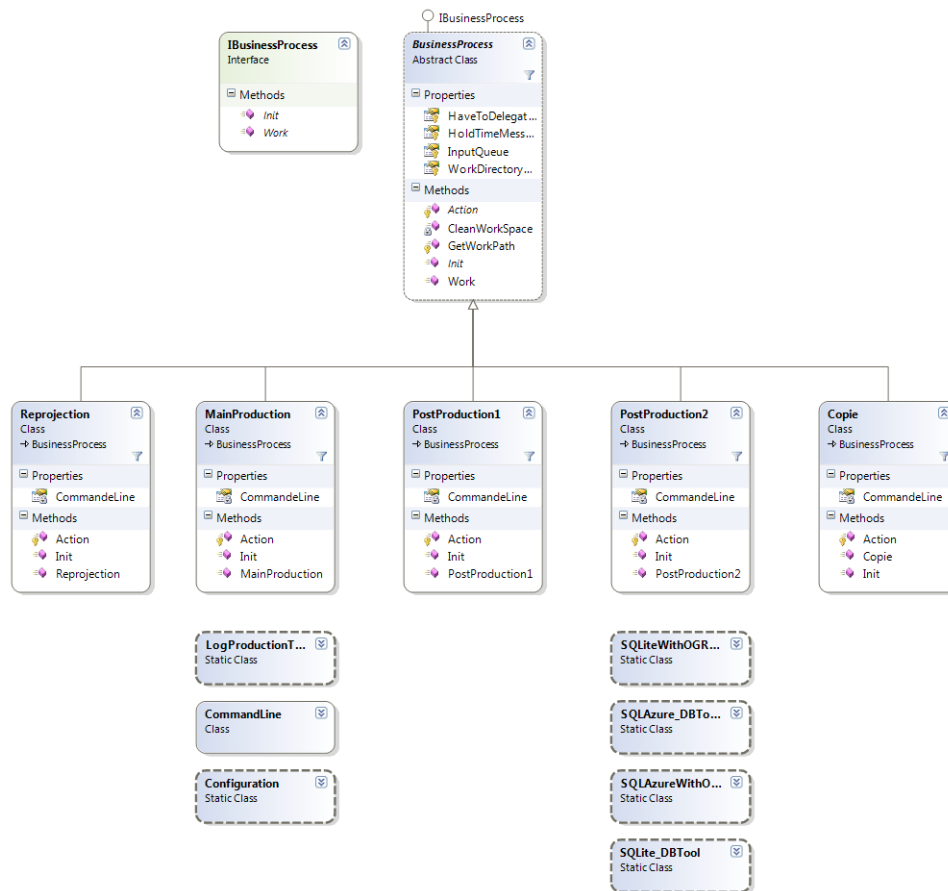
Celles ci se comportent globalement de la même manière. Lors de son instantiation ou de l'appel à *Init()* un objet héritant de *BusinessProcess* met en place son environnement de travail, et crée notamment des chaines de connexion. Ensuite lors de l'appel à la méthode *Work* il vérifie si il a une tâche à faire. La plupart du temps la liste des tâche est implémenté grace à un objet *CloudQueue* de l'API Azure.

Si une tâche est trouvée alors la commande est interprétée à l'aide de la classe *CommandLine* et les fichiers nécessaires sont ramenés des blobs Azure sur la VM (en local) afin de travaillé avec.

Une fois cette étape de mise en place terminée, le code métier est appelé. Il écrit soit directement les résultats dans une table SQL Azure, soit dans un fichier qui sera ensuite remis dans un blob Azure. Enfin le message correspondant à la tâche est supprimé.



FIGURE 3 – Diagramme des classes héritant de BusinessProduction.

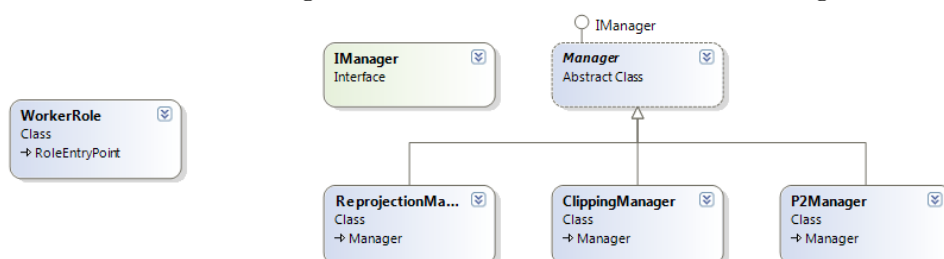


#### 3.4.4 La classe *Manager*

Un autre Worker Role nommé *Manager* sert à générer les milliers de messages envoyés aux queues de travail. C'est un processus similaire au premier.

La classe *WorkerRole* sert de point d'entrée, mais cette fois il n'y a pas de thread de travail et de thread de surveillance. L'objetinstanciera directement un tableau de d'objet héritant de *Manager* et implémentant l'interface *IManager*. Il boucle ensuite sur ce tableau pour savoir si il existe des taches à effectuer. Ces taches sont différentes des précédente puisqu'il s'agit de créer des messages et de remplir des queues.

FIGURE 4 – Diagramme de classe du Worker Role Manager.



### 3.5 Détail des tâches métiers

Nous allons ici aborder fonctionnalités métier.

Elles se séparent en 3 catégories. Les premières peuvent s'exécuter en parallèle par carte, les secondes par RCL<sup>16</sup>. Pour la dernière catégorie le parallélisme est moins évident à mettre en place.

#### 3.5.1 La classe *Reprojection*

Message d'entrée : `-inputcontainer :<input_container_name> -inputfile :<input_file_name> [-outputContainer :<output_container_name >] [-outputfile :<output_file_name>] [-fullProduction : {true; false}] -serveur :<server_name> -dataBase :<data_base_name> -user :<user_name> -pwd :<password> [-overwrite :{true; false}] [-update :{true; false}] [-refFournisseur :<int_provider>`

Au départ d'une production nous possédons des fichiers .s57 dans un conteneur et des tables dans une base de donnée SQL azure.

Le fichier .s57 est ouvert avec l'API Ogr. Chaque lignes de la base de donnée est ensuite recopié dans un fichier sqlite en reprojétant chaque coordonnée spatiale dans un autre référentiel spatial. Les tables OBJECT\_TYPE, PROVIDER\_REFERENCE, CHART\_FID, CHART et S57-OBJECT sont créés et les trois premières remplies. Le résultat est stocké dans un blob.

#### 3.5.2 La classe *MainProduction*

Message d'entrée : `-inputcontainer :<input_container_name> -inputfile :<input_file_name> [-outputContainer :<output_container_name >] [-outputfile :<output_file_name>] -serveur :<server_name> -dataBase :<data_base_name> -user :<user_name> -pwd :<password> [-overwrite :{true; false}] [-update :{true; false}] [-jobToDo :clipping ;trinagulation ;display3d] [-layers :all ;<layers_names >]`

Le fichier précédent est modifié par cette classe. Les géométries des objets sont découpée, on parlera par la suite de clipping, selon la table CATALOGUE\_RCL\_ALL et placée dans la table CLIPPING\_OBJECT. Les géométries de type polygones obtenues sont triangulées en utilisant les bibliothèques j3d<sup>17</sup> et stocké dans la table TRIANGULATION\_2D. Enfin on calcul un point d'affichage pour chaque géométrie, un sorte de barycentre pour l'affichage. Ce résultat est distribué dans les tables DISPLAY\_3D, DISPLAY\_3D\_POLYgone, DISPLAY\_3D\_LINE.

#### 3.5.3 La classe *PostProduction1*

Message d'entrée : `-inputcontainer :<input_container_name> -inputfile :<input_file_name> [-outputContainer :<output_container_name >] [-outputfile :<output_file_name>] [-fullProduction : {true; false}] -serveur :<server_name> -dataBase :<data_base_name> -user :<user_name> -pwd :<password> [-overwrite :{true; false}] [-update :{true; false}]`

Dans les deux post-production le système de traitement d'une tâche est un peu plus complexe car on fait appel à un exécutable, le *displayStyler*, qui est une version modifiée pour l'occasion, d'un code déjà existant et fonctionnel de l'entreprise. L'intégration de ce code directement dans le code produit durant le stage pose des problèmes de compatibilité de framework Microsoft.

16. Row Column Level, Ligne Colonne Niveau, méthode d'agencement en "quadtree pyramid"

17. Java 3d, dll wrapée

C'est pourquoi le choix de modifier les sources de l'exécutable pour qu'il manipule lui même les objets azure a été fait.

La première post-production travail encore par carte. Elle remplit les tables CHART et S57-OBJECTS avec des binaires représentant les données des objets découpés.

### 3.5.4 La classe *PostProduction2*

Message d'entrée : `-inputcontainer :<input_container_name> -inputfile :<input_file_name> [-outputContainer :<output_container_name >] [-outputfile :<output_file_name>] [-fullProduction : {true; false}] -serveur :<server_name> -dataBase :<data_base_name> -user :<user_name> -pwd :<password> [-overwrite :{true; false}] [-update :{true; false}] -refFournisseur :<int_provider> [-jobToDo :clipping;trinagulation;display3d] [-layers :all;<layers_names > ]`

La seconde post-production travaille par RCL, ce qui oblige à télécharger toutes les cartes sqlite du RCL. Une petite particularité afin d'alléger les tables SQL Azure, ici les des binaires générés sont stockés dans des blob azure et leur référence sont insérés dans les tables RCL et S52-RCL. Ces deux tables ne sont plus dans un fichier sqlite mais bien sur SQL Azure afin de pouvoir gérer le parallélisme à gros grain facilement.

### 3.5.5 La classe *Copie*

Message d'entrée : `-inputcontainer :<input_container_name> -inputfile :<input_file_name> [-outputContainer :<output_container_name >] [-outputfile :<output_file_name>] [-fullProduction : {true; false}] -serveur :<server_name> -dataBase :<data_base_name> -user :<user_name> -pwd :<password> [-overwrite :{true; false}] [-update :{true; false}] -refFournisseur :<int_provider> [-jobToDo :clipping;trinagulation;display3d] [-layers :all;<layers_names > ]`

La dernière étape n'est pas parallélisé. Elle s'appuie encore sur le *DisplayStyler*. Elle consiste grossièrement à regrouper tous les résultats précédent en un seul fichier.

## 3.6 Résultats

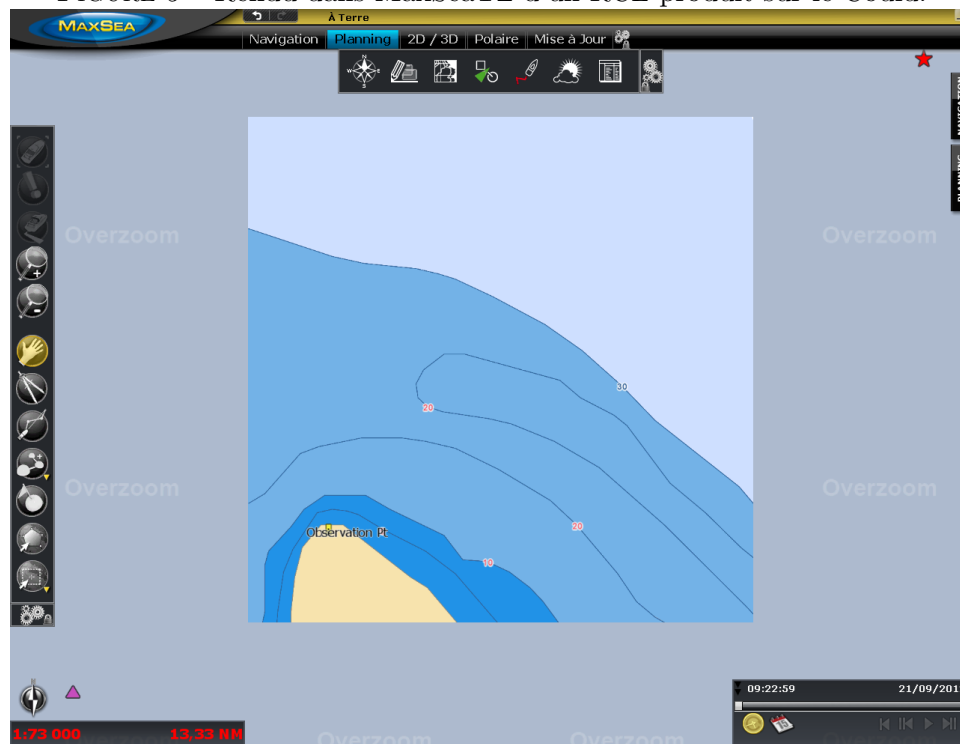
### 3.6.1 Fichier au format dbv

Un fichier dbv créer par la nouvelle chaine de production a put être lu par le logiciel MaxSea TimeZero. La migration de la chaine de production est donc techniquement possible. Seul bémole à cet essai, par manque de temps un seul rcl a été traité. Même si de nombreuses étapes ont été testées en quantité (jusqu'à la post-production2), un plus gros lots de rcl et de carte pour la copie aurai été souhaitable.

### 3.6.2 Performance de travail

La performance n'as pas été le facteur décisif. Cependant on peu dire qu'on bénéficie avec ce système d'une bonne accélération, ou speed-up, même si l'algo est loin d'être optimal. En effet trop de temps est perdu dans les communications, en particulier dans le téléchargement de fichiers sqlite. On remarque que, comme attendu, le temps de production d'un lot donné est inversement linéaire au nombre de worker role, du moins pour les premières étapes où des tests ont été effectués.

FIGURE 5 – Rendu dans MaxSeaTZ d'un RCL produit sur le Cloud.



### 3.6.3 Budgétaires

La question du coût reste assez floue. En effet on ne peut pas faire de comparatif entre l'ancien système et les coût de la production cloud. Nous ne disposons pas d'une estimation du coût annuel de l'infrastructure dédiée à la production vecteur car ces machines sont utilisées pour de nombreuses fonctions simultanément. De plus seul un essai grandeur nature de bout en bout peut vraiment nous donner le coût d'une production vecteur en environnement cloud.

## 4 Amélioration du prototype

Voici une liste, plus ou moins exhaustive, des améliorations possible voir souhaitable du prototype proposé.

### 4.1 Todo-List

### 4.2 Conclusion